Unit 16: Census and Sampling

PREREQUISITES

Using a random number table or a computer random number generator to select simple random samples was covered in Unit 15, Designing Experiments, and is needed in this section.

The description of the U.S. Census and the discussion of sampling ideas have much to say about where social and economic data come from. This unit therefore forms a bridge between mathematics/statistics and social science.

ACTIVITY DESCRIPTION

Give students an opportunity to explore the 2010 (or most current) Census home page. They can use Google or some other search engine to find the url, which changes periodically. (You can try http://www.census.gov.) This activity is designed to make students aware that they can access information from the U.S. Census.

In the activity, students are asked to find the current U.S. population at the time they logged in to the U.S. Census homepage. If students log in at different times, this number will be different. However, just being aware of the size of the current U.S. population will help students appreciate the enormity of the task involved in conducting a census every ten years. Next, students can focus on their own state – and use U.S. Census data to find its 2010 population and some basic demographic information. Question 3 is more open and asks students to compare demographic information on their state with a neighboring state.

THE VIDEO SOLUTIONS

1. The overall accuracy has improved over the years.

2. Representation to Congress for specific regions of the country is apportioned based on U.S. Census information as are federal funds. So an undercount in a certain area of the country means reduced representation and fewer federal dollars compared to what the area should receive.

3. A sample is chosen in such a way that each individual in the population has an equal chance of being selected for the sample.

4. A sample of 150 pounds of potatoes is taken by selecting 5 buckets of potatoes from various locations in the truck. From this sample, a smaller sample of 40 pounds of potatoes is selected for the cooking test (a hole is punched in these potatoes). Remaining potatoes are inspected for defects. There are many other samples along the way to test for: correct thickness, golden color, proper salt content, and satisfactory bag weight.

UNIT ACTIVITY SOLUTIONS

1. Sample answer: On 1/29/13 at 12:40 p.m. EST the population was 315,248,529. (By 1:00 pm the population was 315,248,599.)

2. a. Sample answer: For Massachusetts the 2010 population was 6,547,629.

b. Sample answer: Percentage of males: 3,166,628/6,547,629 × 100% ≈ 48.4%; Percentage of females: 3,381,001/6,547,629 × 100% ≈ 51.6%.

c. Percentage under 18: 1418923/6547629 × 100% ≈ 21.7%; Percentage of 65 & over: 902,724/6547629 × 100% ≈ 13.8%. A higher percentage of the population was under 18 than was 65 or over.

3. Sample answer comparing Massachusetts and Connecticut: Connecticut's population was only 3,574,097 compared to 6,547,629 for Massachusetts. The percentage of males in CT was slightly higher than in MA, 48.7% compared to 48.4%, respectively. The under 18 population in CT was 22.9%, slightly higher than in MA, which was 21.7%. The 65 & over population is also higher in CT at 14.2% compared to only 13.8% in MA. (Students could also compare housing, race, and more on age.)

EXERCISE SOLUTIONS

1. a. The population is not at all clear. Reasonable populations are all residents in the station's viewing area or all viewers of the 6 o'clock news. However, certain viewers who feel strongly about this question could call their friends and ask them to vote.

b. This is a voluntary response poll. The self-chosen respondents have stronger feelings on the issue than the population as a whole. In the case of gun control, the strong feelings of those opposed to gun control are well known. The poll results will almost certainly overstate the percentage of the general public who oppose the ordinance. In addition, different news stations attract different types of viewers. So this poll reflects the opinions of the viewers of this station and not necessarily the residents of the viewing area. Furthermore, many people do not get their news from television. Non-viewers of the 6 o'clock news are not included.

2. Selecting the sample using Table B from *The Basic Practice of Statistics*

01 Agarwal	**08 Dewald**	15 Hixson	22 Puri
02 Anderson	**09 Fernandez**	16 Klassen	23 Rodriguez
03 Baxter	10 Frank	17 Mihalko	**24 Rubin**
04 Bowman	11 Fuhrmann	18 Moser	25 Santiago
05 Bruvold	12 Goel	19 Naber	26 Shen
06 Casella	13 Gupta	**20 Petrucelli**	27 Shyr
07 Cordero	**14 Hicks**	21 Pliego	28 Sundheim

There are 28 students. Label them 01 to 28 in alphabetical order.

Line 136 of Table B is:

08421 44753 77377 28744 75592 08563 79140 92454

Reading two-digit groups and skipping those not used as labels, our sample contains the students labeled 08 14 20 09 24. These names have been bolded in the list above.

Selecting the sample using Excel's Rand ()

Step 1: Enter the names into column A of an Excel spreadsheet.

Step 2: In column B use Rand () to generate a column of 28 random numbers.

Step 3. Use Data>Sort to order the names in column A by their corresponding random number in column B.

Step 4. Select the first 5 names from the sorted list from Step 3.

Name	Rand
Agarwal	0.47616
Anderson	0.42692
Baxter	0.44405
Bowman	0.27579
Bruvold	0.12247
Casella	0.82995
Cordero	0.12318
Dewald	0.57609
Fernandez	0.31248
Frank	0.62985
Fuhrmann	0.93206
Goel	0.80434
Gupta	0.44848
Hicks	0.77278
Hixson	0.64893
Klassen	0.84144
Mihalko	0.19905
Naber	0.06491
Petrucelli	0.3356
Pliego	0.43135
Puri	0.42294
Rodriguez	0.963
Rubin	0.3613
Santiago	0.45452
Shen	0.13584
Shyr	0.54541
Sundheim	0.03402

Names ordered by Rand
Sundheim
Naber
Bruvold
Cordero
Shen
Mihalko
Bowman
Fernandez
Petrucelli
Rubin
Puri
Anderson
Pliego
Baxter
Gupta
Santiago
Agarwal
Shyr
Dewald
Frank
Hixson
Hicks
Goel
Casella
Klassen
Fuhrmann
Rodriguez

Sample answer: Sundheim, Naber, Bruvold, Cordero, Shen

3. a. The population would be all the Hudson Valley Patch Facebook readers or it could be all residents of the Hudson Valley region in New York state. (If the latter, the sample in (b) will miss all of the non-Facebook users in Hudson Valley.)

b. The sample would be the readers who voluntarily voted for the worst Valentine's Day gift.

c. No. First, not all Hudson Valley residents are on Facebook and connected to Hudson Valley Patch. In particular, the votes do not represent the opinions of non-Facebook users.

4. a. Population: all home sales in Worcester County, Massachusetts; sample: 50 home sales.

b. Population: all veterans who served in combat; sample: the 25 veterans examined by the psychologist.

c. Population: all seniors attending Eastern Connecticut State University; sample: 20 seniors questioned by the educator.

Some students may decide that the population is all students attending Eastern. However, then the educator has selected an unrepresentative sample because it only contains seniors.

REVIEW QUESTIONS SOLUTIONS

1. a. This question can be done using Table B from *The Basic Practice of Statistics*, but it is very tedious to do so. The sample answer relies on use of Minitab's Uniform random number generator.

Sample answer: We used Minitab's Uniform random number generator to assign a random number between 0 and 1 to each of the 48 students. Then we sorted the students by arranging their assigned random numbers from smallest to largest. The first 8 students in the sorted list were selected for the first group. (See sorted list in solutions to (b).)

b. Sample answer: The first 8 students in the ordered list are assigned to Section 1, the second set of 8 students are assigned to Section 2, and so forth. A complete listing of students and their sections appears below.

Names	Uniform	Section	Names	Uniform	Section
Juarez	0.0026	1	Elsevier	0.5289	4
Swokowski	0.0083	1	Stevenson	0.5366	4
Scott	0.0762	1	Fernandez	0.5616	4
Burns	0.1651	1	Barrett	0.5665	4
Schiller	0.1810	1	Poe	0.6165	4
Erskine	0.1882	1	Beerbohm	0.6184	4
Hyde	0.2005	1	Garcia	0.6562	4
Flury	0.2192	1	Rodriguez	0.6752	4
Taylor	0.2227	2	Orsini	0.6785	5
Arnold	0.2618	2	Putnam	0.6896	5
Jones	0.2908	2	Rowley	0.6905	5
Perlman	0.3187	2	Deneuve	0.7393	5
Nguyen	0.3546	2	Neale	0.7754	5
Kempthorne	0.3601	2	Moore	0.7772	5
Chang	0.3770	2	Campbell	0.7804	5
Quincy	0.4251	2	Dodington	0.7896	5
Prizzi	0.4559	3	Drummond	0.7964	6
Smith	0.4751	3	Oakley	0.7978	6
Ward	0.4786	3	Levine	0.8051	6
Hardy	0.4828	3	Ashford	0.8823	6
Colon	0.4888	3	Bartkowski	0.8961	6
Munroe	0.4966	3	Martinez	0.8965	6
Holmes	0.5011	3	Randall	0.9285	6
Vuong	0.5087	3	Rostenkowski	0.9408	6

2. a. The population is the set of students entering a college. The sample is the group of students questioned by this professor during their orientation.

b. The population consists of patients suffering from arthritic knees. The sample consists of 10 of the physical therapist's patients who had arthritic knees.

3. a. The population consists of all students who graduated from this university at least five years ago. (That way you can determine what they were doing 5 years after graduation.) You may want to narrow the population to students who graduated between 5 and 8 years ago or narrow even further to students who graduated exactly 5 years ago.

b. The cost to conduct a census would be too high and it would take too long to gather the results. If a survey is mailed to the graduates, you would have to track down those who didn't respond and try to get their information with one-on-one phone calls or home visits. So, this would greatly add to the time required to gather this information. Furthermore, no matter how hard you tried, it would be impossible to track down every graduate who graduated 5 or more years ago. Some will have left the area (or even the country) without providing forwarding addresses. Given a complete list of graduates in the population of interest, you could focus on a sample. Since the size is small, you could contact each person in the sample.

4. a. Sample answer:

Pros of conducting a census:

- If it is possible to contact everyone in the population, you get a true measure of the proportion of the population that supports the measure.
- Not only do you know the overall population proportion supporting the measure, you can also determine if there are specific subgroups of the population (even if the subgroup is a low percentage of the population) that oppose the measure.

Cons of conducting a census:

- It may not be possible to contact everyone in the population in one month. Some people may be away that month. Others may not want to be contacted at all and hence those people's views will not be represented in the census.
- If it is a large population, you may not have the manpower to contact everyone in one month.
- It will cost more to conduct a census than to take a sample.

b. Sample answer:

Pros of taking a sample:

- Costs would generally be lower than for a census.
- If good sampling techniques are used, the results collected from the sample should be representative of the views of the population.
- It may take less time to gather and analyze the data.

Cons of taking a sample:

- Data may not be representative of the population. This is particularly true if the sample size is small or if an inadequate sampling plan (such as voluntary sampling) is used.
- There is variability due to sampling. Different samples could lead to different results.
- Since you are working with a sample, you may not be able to get detailed information about certain subgroups within the population who oppose the measure, particularly if those subgroups are small in comparison to the population. (This problem may be fixed by revising the sampling plan.)